Open in app ↗ Sign up Sign in

Medium Search

# The Three-Room Problem: How AI-Assisted Analysis Challenges Traditional Chip Security (And Why the Original Maker Might End Up Paying Licensing Fees)

15 min read · Just now

Berend Watchus Follow

Listen Share

Author: Berend Watchus December 07, 2025 Follow up article to the previous artice: ‘Amazon’s $10 Billion Bet That Nobody Can Copy Its AI Chips’

# The Three-Room Problem: How AI-Assisted Analysis Challenges Traditional Chip Security (And Why the Original Maker Might End Up Paying Licensing Fees)

**Author: Berend Watchus**

**Date: December 2025**

**Category: Cybersecurity, AI Security, Hardware Security**

## Executive Summary

Amazon's Trainium3 chip relies on security through complexity, with industry analysts claiming reverse-engineering would require 4–8 years and $1–3 billion. **Amazon's strategic choice to not file patents** (as revealed in "Amazon's $10 Billion Bet That Nobody Can Copy Its AI Chips")

**Amazon's $10 Billion Bet That Nobody Can Copy Its AI Chips**

Author: Berend Watchus December 4, 2025

medium.com

represents false security — trusting that nobody can reverse-engineer fast enough and cheap enough to out-compete them in the market. This article demonstrates why those assumptions are obsolete.

**Key findings:**

- **The methodology is published:** The AKA (Autonomous Knowledge Accelerator) approach — documented by an independent researcher with no formal engineering education — enables systematic cross-domain technical analysis using commercial conversational AI
- **Tier 5 capability (lowest tier):** €1,000/month enables 50 automated LLM agents

running 24/7 parallel analysis — accessible to any small business or motivated individual

- **Tier 1 capability at budget cost:** A three-room €2M operation (Technology R&D and market research firm) appears 100% legitimate to any inspection while delivering Tier 1 nation-state-level performance. Not spending impressive amounts — just €2M — but wielding Tier 1 analytical powers through methodology optimization
- **Everything is legal:** From day one through market launch, including selling nearly identical chips — zero hiding from authorities, only from competitors
- **The timeline compresses:** Projected 6–9 months versus Amazon's assumed 4–8 years
- **The ultimate irony:** Competitors could patent specific implementations and force Amazon to pay licensing fees for their own architecture — which is not threatening, not extorting, not industrial espionage. It's simply being the first party to claim and protect an implementation with a patent. Completely legal competitive strategy.

**Critical context:** This scenario overlaps 99% with illegal industrial espionage in patent-heavy industries. However, Amazon's "security through secrecy, no patents" strategy makes this entirely legal. The team wouldn't even need to produce or sell anything — they could simply reverse-engineer, rush to patent specific implementations, then sue Amazon for infringement of their own architecture. **This is worse than patent trolling:** You're not suing adjacent overlapping chip companies — you're 100% targeting the original mega-investor, R&D specialist, and original manufacturer/seller. You can weaponize their own market share and track record against them in damages calculations: "How much is this patented implementation worth? Well, Amazon sold X million chips using it, at $Y each…" The absence of patent protection Amazon relies on for security is precisely what enables both legal competitive reverse-engineering and subsequent patent weaponization targeting the original creator.

**Target audience:** Security professionals, IT R&D executives, market analysts,

government agencies concerned with emerging threats to intellectual property protection models.

— — — — — — — — — — — — — — — — — — —

## Introduction: When Security Assumptions Meet New Capabilities

> **"It is not like we are talking about patented stuff, we are talking about the opposite — competition hoping you would NOT do this and could not do this and could not pay for it and get away with making profit with the process."**

Amazon's recent Trainium3 chip announcement came with confident assertions about security through complexity. According to industry analysts, reverse-engineering such advanced chips would require 4–8 years and $1–3 billion in resources. The underlying assumption: only nation-states or the largest competitors possess the expertise and resources for such efforts.

But what if those assumptions are already obsolete?

As an independent AI researcher who has spent 2024–2025 exploring AI-assisted analysis methodologies, I've documented something concerning: the barrier to complex technical analysis has dropped dramatically. Not theoretically — demonstrably.

## Who I Am: Why This Background Matters

Why talk about me? Well, that becomes relevant very quickly when you learn who I am **not**.

I am a high school dropout who both today and in the 90's would not pass basic physics, chemistry, math, or IT tests. I cannot program. By traditional metrics, I should not be able to contribute to cutting-edge technical fields.

Yet I am the world's most published and founding researcher on mirror testing methodologies for artificial intelligence. My work has established the foundational frameworks that define this emerging subfield — adapting classical mirror tests (the rouge test for human infants and Gallup mirror test for primates) into comprehensive AI mirror testing protocols covering all variations: physical

embodiment, virtual environments, emergent versus programmed self-awareness, and cross-platform identity persistence.

This research represents approximately 5% of my total body of work through November 2025. All papers have passed rigorous scientific screening by the Preprints.org editorial board (which maintains a ~50% rejection rate). The 2024 foundational work was compiled into the book "AI and Mirror testing, Science Papers 2024" (ISBN: 9789465200927), with continued AI mirror test publications throughout 2025 on implementation specifics, long-term AI mirror test research protocols, and accessibility approaches for upgrading older robots with modern self-recognition capabilities.

My broader research encompasses AI theory, consciousness frameworks, embodiment principles, and the development of systematic methodologies for cross-domain technical analysis — **all without formal education in any of these domains**.

This matters because if someone with my educational background can achieve these outputs, the barrier to complex technical analysis has fundamentally changed. That's what this article is about.

## Background: The AKA (Autonomous Knowledge Accelerator) Methodology

In 2024, I published a series of AI theory papers on preprints.org creating new concepts, frameworks, and metrics for intelligence in general — applying substrate-agnostic models across animals, humans, AI, NPCs (embodied AI in simulated environments), robots (embodied AI), and even SETI research. These papers passed editorial screening (which rejects ~50% of submissions) and were later cited by university research teams — and, less flatteringly, plagiarized by others (a case I won).

In 2025, I developed a Theory of Everything (TOE) and the Law of Optimized Complexity (LOOC) — positioned as a twin theory to the second law of thermodynamics. These theoretical frameworks proved key in enabling my autonomous research methodology (finalized by end of 2025) to perform on

demand at specific magnitude-of-improvement scales — allowing requests like "achieve 1,000× to 10,000× efficiency improvement" to produce results within the specified range.

Through this research process, I developed what I call the AKA (Autonomous Knowledge Accelerator) — my specific implementation of autonomous research methodology, a systematic approach for using conversational AI to perform expert-level analysis in domains where you lack formal expertise. AKA runs on commercially available online conversational AI through prompting techniques. The methodology has two components: (1) systematic prompting protocols, and (2) upgrading the LLM within each chat session by having it read my papers and blog articles from the self-referential knowledge graph, effectively expanding its context with specialized frameworks (LOOC, TOE, substrate-agnostic models, etc.). I documented this approach and published it in November 2025.

A key component was building an unconventional self-referential knowledge graph. Without IT skills to create traditional knowledge graphs, I relied on pattern recognition and intuition about information flow and computation. In an analog way, I constructed this knowledge graph through my blog articles and research papers themselves — each publication serving a dual purpose: contributing knowledge to the field while simultaneously becoming nodes in a semantically woven self-referential system. (Other articles explain how these are semantically interconnected.)

The methodology isn't theoretical. I've applied it repeatedly across domains:

**Turbofan Reverse-Engineering (2 hours):**

- Background: Zero physics education
- Tools: Commercial conversational AI
- Output: Complete technical analysis of turbofan engine architecture
- Purpose: Blog article for System Weakness
- Result: Published, technically accurate, readable

The turbofan analysis was casual — just content for a blog post. But it demonstrated something significant: conversational AI can perform complex engineering analysis (read: reverse-engineer!) when guided by systematic methodology, even when the human operator lacks domain expertise.

**Other applications documented:**

- Cross-domain technical synthesis
- Efficiency gains: 200×, 3,700×, 8,700× in various applications

That methodology is now public. Published. Replicable.

## The Floor: What One Person Can Do

Let me establish the baseline — what I've personally demonstrated:

**Investment:** Free conversational AI (2024 to October 2025), then €20/month subscription
**Time:** 2 hours to maximum one day depending on complexity
**Education:** Zero formal training in engineering domains analyzed
**Resources:** Old laptop (2024 to November 2025, couldn't play YouTube above 1080p, Windows 10), internet connection, conversational AI access
**Team:** Solo

**Output:**

- Complete reverse-engineering analyses
- Technical documentation
- Published specifications
- Quantum Watchman security framework (original concept, heavily upgraded using LLM before publishing "IKEA for Spies: The Quantum Watchman Goes Mail-Order")
- LOQSNC architecture article: NIST standardized post-quantum cryptography

algorithms in August 2024, but running them on IoT devices crashes battery life from 5 years to 6 months. The LOQSNC (LOOC-Optimized Quantum-Safe Neural Cryptoprocessor) achieves 8,700× efficiency improvement through thermodynamic system design rather than component optimization, making quantum-safe IoT economically viable. Totally made by my AKA (Autonomous Knowledge Accelerator).

- Validated by readers with domain expertise

This is the floor. Not a hypothetical — a documented track record spanning multiple complex engineering domains.

## The Industrialization Threshold: €1,000/Month

The methodology scales. Here's what becomes possible at modest budget levels:

**€1,000/month gets you:**

- 50+ LLM API instances running in parallel
- Automated agent coordination
- 24/7 exploration of design space
- Systematic cross-domain synthesis
- Cloud compute for orchestration (or for privacy-conscious organizations: on-premise setup with local LLMs, eliminating external dependencies and API logs — but that would cost more)

**Timeline compression:**

- My solo turbofan analysis: 2 hours
- Automated 50-agent system: Continuous operation
- Output: Not one analysis, but systematic exploration across hundreds of variations

**Who has €1,000/month?**

- Any mid-sized company's coffee budget
- Individual researchers with modest grants
- Small consulting firms
- Hobbyists with day jobs

This isn't nation-state territory. This is "any motivated individual with a credit card" territory.

## The Three-Room Problem: A Hypothetical Scenario

Now let's engage in threat modeling. What would a serious, well-funded operation look like?

**Total Budget: €2M**
**Timeline: 6–9 months**
**Location: Any industrial park, any city**

## Room 1: "Competitive Analysis Lab"

**Cover Story:** Technology R&D and market research firm

**Visible Activities:**

- Imaging equipment and microscopes
- Analysis workstations
- Sample processing

**Actual Function: Training Data Acquisition**

This is completely legal competitive analysis:

- Purchase older chips on open market (Intel, AMD, NVIDIA, ARM — all legal)

- Subscribe to TechInsights professional tear-down services (€10k/year, legitimate business)
- Mine patent databases (free, public information)
- Collect academic papers (free, public research)
- Industry conference materials (public presentations)

**Result:** Comprehensive 50-year corpus of chip design evolution

**The Critical Detail:** If authorities walk in at any time — 2pm on a Tuesday, surprise inspection, whatever — everything is completely legitimate. They'd see people literally studying and mapping chips, which is exactly what the business claims to do. Competitive analysis, reverse-engineering for interoperability, technology assessment — all legal. Even if inspectors see Intel chips on workbenches, imaging equipment analyzing die photos, documents mapping transistor layouts, the response is simple: "We do technology consulting. This is what we do for clients. All purchased hardware, all public data, nothing classified."

**And critically: Not just the analysis is legal — using the reverse-engineered ideas to enter the market is 100% legal too.** From day one to the last day of selling chip designs or manufacturing nearly identical chips with copied EVERYTHING (avoiding only minor specific patented details if any) — completely legal. Reverse-engineering for competitive purposes, learning from competitor architectures, and implementing similar designs are all protected competitive activities. **There's zero hiding from the law or authorities — it's only hiding from competitors.** The operation is completely transparent to legal inspection while invisible to market surveillance.

**Even more devastating:** Competitors could reverse-engineer the designs, patent specific implementations or improvements, and then force the original company to either pay licensing fees or stop using their own architecture in certain jurisdictions. This isn't theoretical — it's standard competitive patent strategy.

**Every aspect passes inspection:**

- Administration: Normal business operations, proper registration, tax compliance
- Finance: Legitimate expenses (chip purchases, equipment, staff salaries, subscriptions)
- Logistics: Legal procurement of components, standard vendor relationships
- Daily activities: Technology analysis, research, documentation — all legal
- Growing knowledge base: Publicly available data, purchased materials, original analysis
- Database: Compiled from legal sources (patents, papers, purchased chips)

No laws broken. No theft. No hacking. No espionage. **This isn't about patents or trade secrets — it's the opposite. It's about what the competition hopes you WON'T do, assumes you can't afford to do, and doesn't believe you can profit from.** The cover story isn't a lie — it's the literal truth presented without mentioning why they're doing it.

**Annual Cost:** €50–100k
**Legal Status:** 100% legitimate business activity
**Visibility to Targets:** Zero

### Room 2: "Analysis Team Offices"

**Cover Story:** Competitive analysis consultants

**Visible Activities:**

- Normal office environment
- Workstations, monitors, whiteboards
- Looks like any tech startup

**Actual Function: Expert Validation Hub**

**Team Composition:**

- 1 Expert/Leader: PhD, 10+ years chip/materials experience (€500k for 6–9 months)
- 2 Specialists: One chip design, one AI/ML (€700k total for 6–9 months)
- Support staff: Security, admin, IT (€100k)

**Activities:**

- Review outputs from Room 3's automated systems
- Validate technical feasibility
- Direct exploration priorities
- Apply human expertise to AI discoveries
- Iterative refinement

**Annual Cost:** €1.2M (personnel)
**Legal Status:** Competitive analysis consulting (legal)
**Visibility to Targets:** Zero

## Room 3: "Server Infrastructure"

**Cover Story:** Data processing and analysis systems

**Visible Activities:**

- Climate-controlled server room
- Standard enterprise IT setup
- Every company has one

**Actual Function: The AI Analysis Factory**

Get Berend Watchus's stories in your inbox

Join Medium for free to get updates from this writer.

Enter your email

Subscribe

**Infrastructure:**

- €100k one-time: High-end servers, GPUs, storage
- 50–100 local LLM instances (Llama, Mistral, Qwen — all open-source, free)
- Fine-tuned on 50-year chip design corpus from Room 1
- Running AKA (Autonomous Knowledge Accelerator) methodology at industrial scale
- Exploring cross-domain synthesis 24/7

**Key Advantages:**

- Air-gapped (no cloud, no external monitoring)
- Unlimited queries (no rate limits, no ToS restrictions)
- Zero API logs (completely invisible to platform monitoring)
- On-premise (full control, complete privacy)

**Operating Cost:** €50k/year (power, cooling, maintenance)
**Legal Status:** Private research infrastructure (legal)
**Visibility to Targets:** Absolute zero

## The Workflow

**Months 1–2: Setup**

- Secure facility
- Deploy infrastructure
- Recruit team
- Begin training data acquisition

**Months 3–4: Training Phase**

- Feed 50 years of chip designs to LLMs
- Fine-tune on pattern recognition
- Implement cross-domain synthesis methodology
- LLMs learn: “This is how chips evolved 1970–2025”

**Months 5–7: Intensive Synthesis**

- 50–100 LLMs exploring design space 24/7
- Generating 100+ novel architecture variations daily
- Expert team validates promising outputs
- Iterative refinement based on feasibility analysis

**Months 8–9: Deliverable Production**

- Complete alternative chip architecture
- Manufacturing pathway documentation
- Cost analysis
- Patent landscape review

- Competitive positioning strategy

## The Deliverable

**What €2M Buys in 9 Months:**

✓ Novel chip architecture (potentially superior to target)
✓ Complete technical specifications
✓ Manufacturing feasibility study
✓ Cost-benefit analysis
✓ Patent freedom-to-operate analysis
✓ Zero attribution (air-gapped, no cloud trail)
✓ Plausible deniability (legitimate research front)

## The Threat Actor Landscape

**Who has €2M and motivation?**

**Corporate Actors:**

- Any F500 tech company (quarterly R&D rounding error)
- Mid-sized chip competitors (strategic investment)
- Private equity firms (due diligence for acquisition)
- Consulting firms (client-funded analysis)

**State-Adjacent Actors:**

- Small country intelligence budgets
- University “research” fronts
- State-owned enterprises
- Defense contractors

**Others:**

- Well-funded startups (investor-backed competitive analysis)

- Industry consortiums (pooled resources)
- Motivated individuals with significant capital

**Even Criminal Enterprises:**

- IP theft syndicates (high-value targets)
- Corporate espionage operations
- Technology acquisition for resale

The critical insight: €2M is not nation-state money. It's accessible to hundreds of thousands of entities globally.

## The Security Model Gap

**Amazon's Assumed Defense:**

- Attack cost: $1–3 billion
- Attack timeline: 4–8 years
- Required expertise: Elite, scarce specialists
- Adversary profile: Nation-states only

**Projected performance and timeline in the scenario:**

- Attack cost: €2M total
- Attack timeline: 6–9 months
- Required expertise: Published methodology + team coordination
- Adversary profile: Anyone with small business loan

**The Math:**

- Cost difference: 500–1,500× less than assumed

- Time difference: 5–10× faster than assumed
- Expertise difference: Methodology is published, replicable

## Why This Matters: The Undetectable Threat

Traditional security monitoring assumes:

- Cloud-based attacks (API logs exist)
- Network intrusions (traffic analysis possible)
- Insider threats (behavioral monitoring works)
- Physical theft (security cameras, access logs)

The three-room scenario is none of these:

- ✗ No cloud footprint (on-premise LLMs)
- ✗ No network intrusion (uses public data)
- ✗ No insider access needed (external analysis)
- ✗ No theft (competitive intelligence)

**It's completely legal. Entirely invisible. Totally undetectable.**

The operation looks identical to thousands of legitimate competitive analysis firms. There's no signature to detect, no behavior to flag, no anomaly to investigate.

## The Training Data Advantage

Here's what makes this particularly challenging to defend against:

**All of this is public information:**

- 50+ years of chip designs (purchasable)
- Thousands of patents (free databases)
- Academic papers (public research)

- Industry conferences (published presentations)
- Professional tear-down reports (commercial services)

**Legal acquisition methods:**

- Buy chips on open market
- Subscribe to analysis services
- Mine patent databases
- Collect academic literature
- Industry partnerships

The LLMs in Room 3 aren't just analyzing Trainium3. They're trained on the entire evolutionary history of chip design. They understand:

- What patterns led from 1970s processors to 2025 architectures
- How design solutions evolved across generations
- What trade-offs define each era
- Where the technology trajectory points next

**They don't reverse-engineer. They predict.**

Because they've seen every step of the evolution that led to the current generation.

## The Methodology Is Already Public

This isn't hypothetical threat modeling. The methodology exists and is published:

- November 19, 2025: AKA (Autonomous Knowledge Accelerator) methodology paper released
- Track record: Turbofan analysis (2 hours), security frameworks, cross-domain synthesis

- Efficiency gains documented: 200×, 3,700×, 8,700× across applications
- Replication instructions: Available to anyone

The floor is established. The approach is validated. The barriers have fallen.

## Implications for Hardware Security

**For Chip Manufacturers:**

The security-through-complexity model needs reevaluation. If the barrier to analysis has dropped from "nation-state resources" to "motivated individual with published methodology," then:

1. **Assume analysis is occurring:** Multiple entities are likely analyzing your chips right now
2. **Timeline compression:** Competitive intelligence cycles are measured in months, not years
3. **Democratized capability:** The threat isn't just major competitors anymore
4. **Public data vulnerability:** Your evolutionary trajectory is visible in patent filings

**For Security Professionals:**

Traditional threat models may not account for:

- AI-assisted analysis capabilities (new, rapidly evolving)
- Methodology democratization (published, replicable)
- Cost barrier collapse (€2M vs. $3B assumed)
- Undetectable operations (air-gapped, legal, invisible)

## The Questions We Should Ask

**How many three-room operations exist right now?**

You can't detect them (air-gapped, no cloud trail). You can't prosecute them (all legal,

public data). You can't prevent them (methodology is published).

**What's the actual timeline for chip analysis in 2025?**

My turbofan analysis: 2 hours, zero physics knowledge. Professional team with 50 LLMs: [?]

**Who are the actual threat actors?**

Not just nation-states. Not just major competitors. Anyone with:

- €2M budget (accessible to thousands of entities)
- 6–9 months timeline (manageable project scope)
- Published methodology (free, replicable)

**Is security-through-complexity still viable?**

When analysis scales from "elite experts" to "automated LLM swarms trained on public data," complexity might not protect as assumed.

## Conclusion: The Paradigm Has Shifted

I'm not claiming chip security is broken. I'm observing that the assumptions underlying current security models may not reflect 2025 capabilities.

The turbofan article took me 2 hours. I don't know physics. The LLM did the engineering work. That's the documented floor.

The ceiling — what well-funded, professionally-staffed operations can achieve — is constrained primarily by:

- Budget (€2M is accessible)
- Timeline (6–9 months is manageable)
- Methodology (published, replicable)

**The wheel was already invented.** Competitive analysis firms, intelligence agencies,

and corporate R&D teams have been analyzing competitor technologies for decades.

**What changed:** AI-assisted analysis dropped the expertise barrier while published methodologies democratized the approach.

**What hasn't changed:** It's still legal. Still invisible. Still based on public information.

The three-room problem isn't whether such operations can exist. It's that we can't know how many already do.

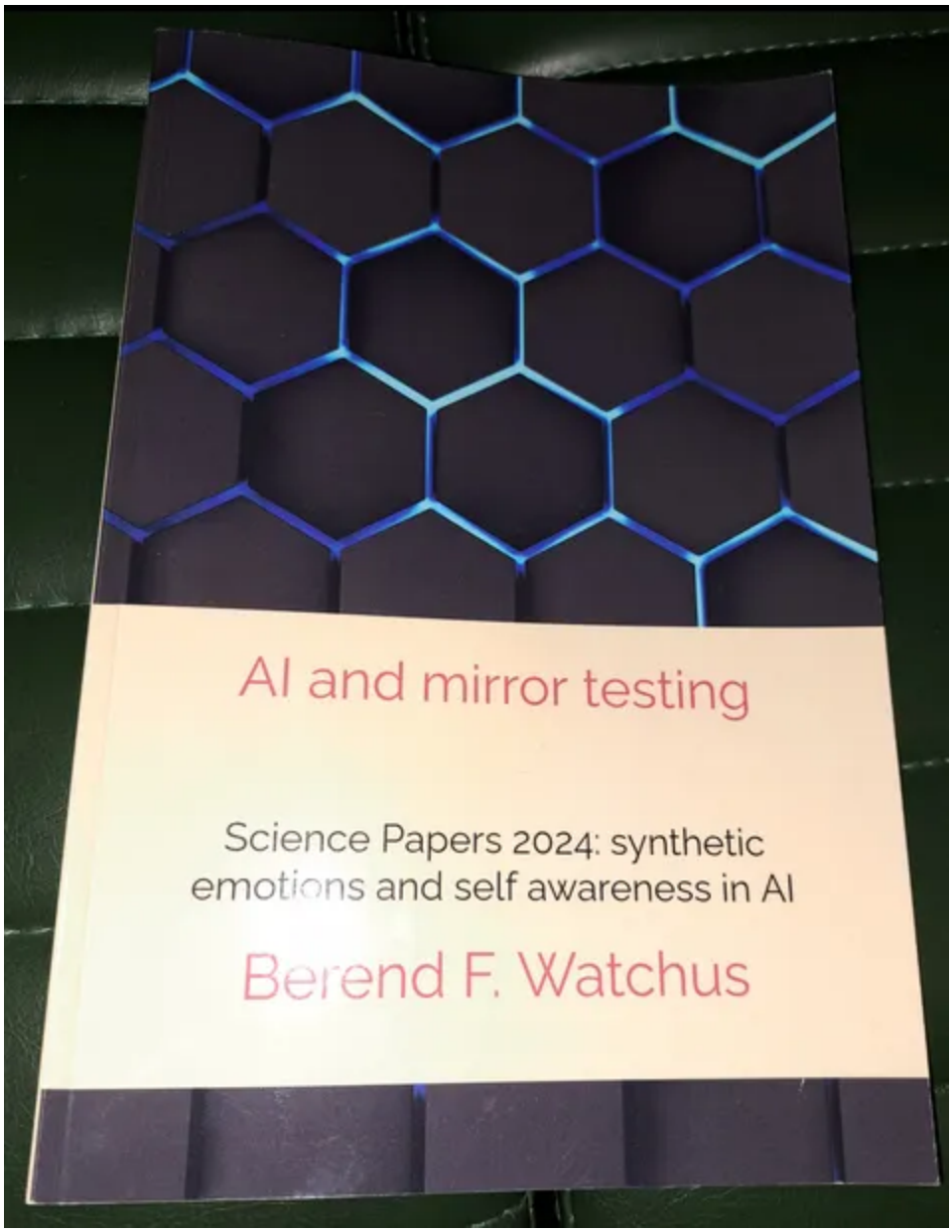
## About the Author

Berend Watchus is an independent AI researcher and security analyst. His work focuses on AI capabilities, security frameworks, and cross-domain methodology development. Published papers include research on AI consciousness, embodiment theory, and the UPRA Framework for unintended data persistence in LLMs. The AKA (Autonomous Knowledge Accelerator) methodology was published in November 2025.

**Berend Watchus's publications**

SciProfiles provides open access to the scholarly community, offering comprehensive profiles where researchers can...

sciprofiles.com




https://archive.org/details/20251113-183440/20251113_183412.jpg

Notable publications include: The CDCL Framework: Unveiling the Hidden Threat Landscape of AI Deception and Control (2025), ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies for Games and Industry Simulations (2025), Beyond the Imitation Game: The Inadequacy of the Turing Test for Modern AI (2025), From Chaos to Efficient Computing: Universal Compression, the P vs NP Divide, and the Hidden Hand of Code (2025), The Unified Model of Consciousness: Interface and Feedback Loop as the Core of Sentience (2024), and Linguistic Trojan Horse: Why LLM Steganography Just Broke AI Safety (2025) — a review of “LLMs can hide text in

other text of the same length" by Antonio Norelli and Michael Bronstein (published October 24, 2025) analyzing the implications and creating scenarios for his breakthrough finding.

## Disclaimer

This article presents threat modeling scenarios based on publicly available information and published methodologies. No actual operations are described. All scenarios are hypothetical explorations of potential vulnerabilities in current security models. This work is intended to inform security professionals about emerging capability gaps, not to encourage or enable malicious activity and not to execute the hyper-competitive but legal scenarios. They are meant for cybersecurity audiences and informative to IT R&D professionals and executives. May also be interesting for market analysts and government agencies.

**Critical Context:** Besides the learning directly from this case and scenario, it overlaps 99% with actual illegal industrial espionage — in cases where there are many detailed patents for tech concepts, as is the case in most products. However (see referenced Amazon article), this is just not the case with Amazon and their chip. Their security strategy of "we are safe, we are not even filing patents" leaves them vulnerable, even though they put their products out there in the market for everybody to see, buy, and study. The very absence of patent protection that Amazon relies on for secrecy makes this entire scenario completely legal, while the same operational structure and methodology would constitute espionage if applied to patent-protected technologies.

Industrial Espionage | Research And Development | Infosec | Cybersecurity | Computer Science

Follow

## Written by Berend Watchus

25 followers · 44 following

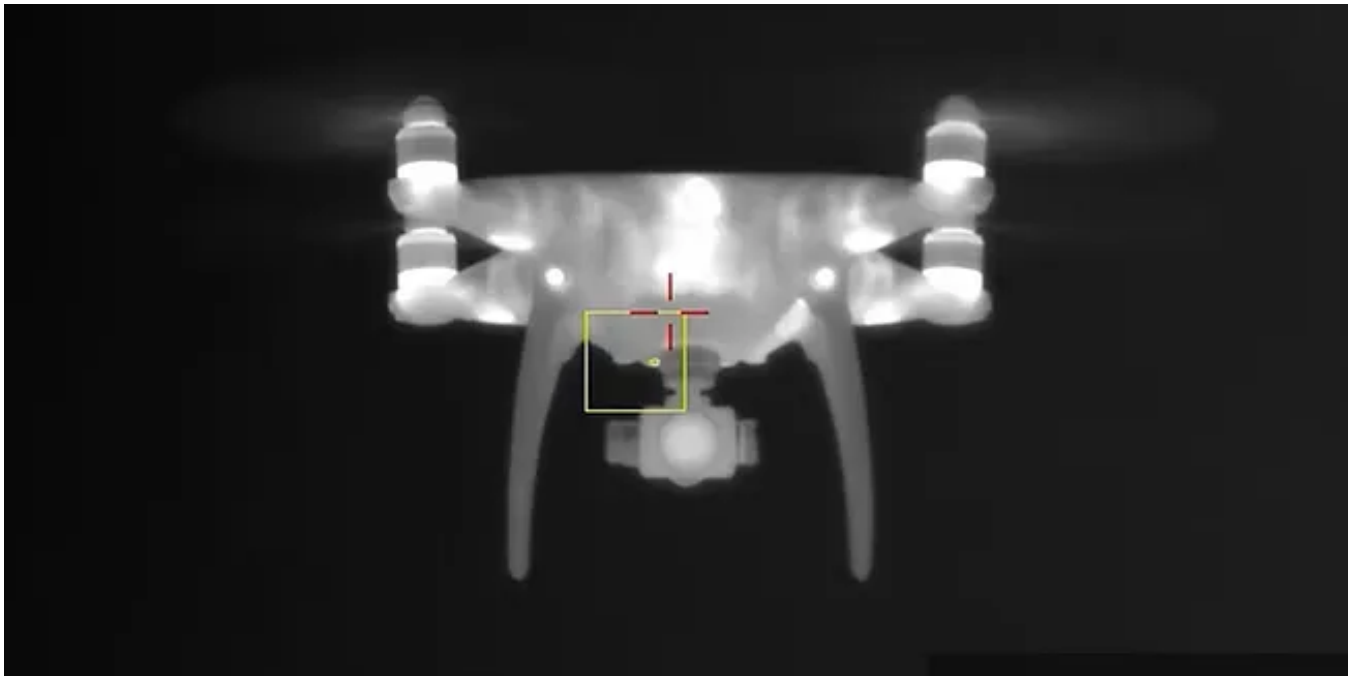
Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer

## No responses yet

Write a response

What are your thoughts?

## More from Berend Watchus

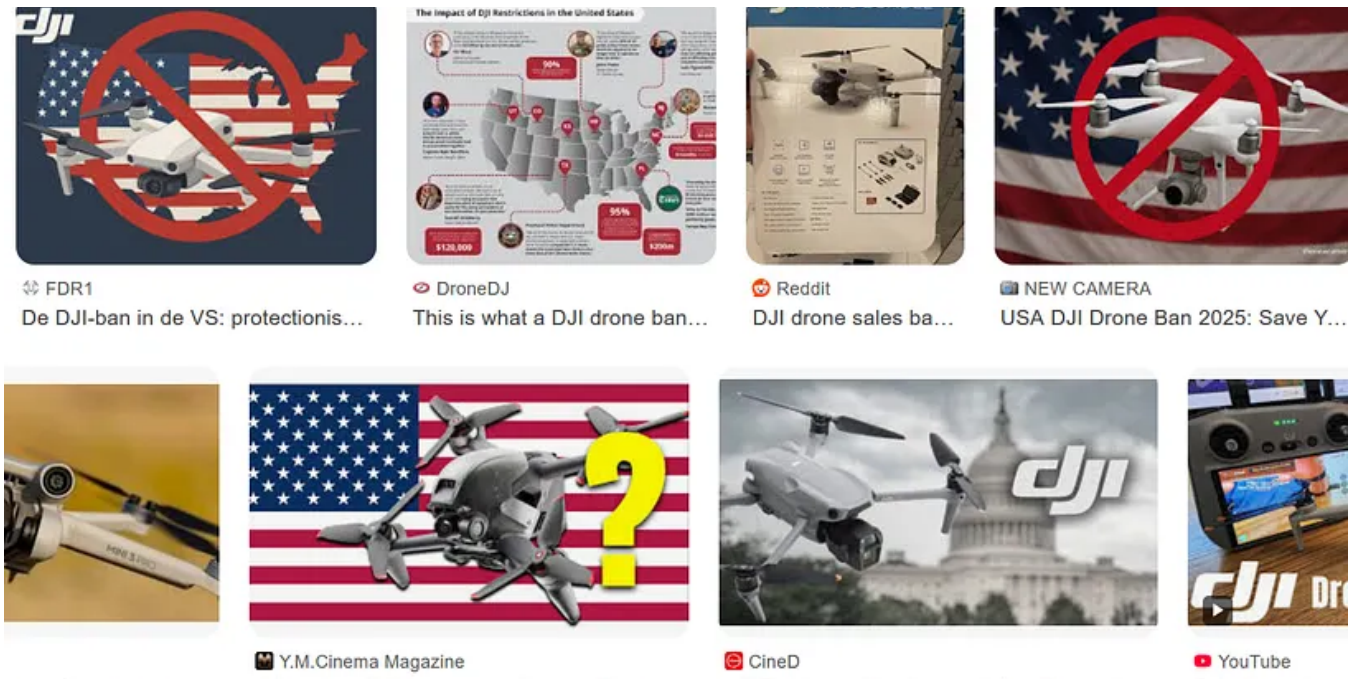


In System Weakness by Berend Watchus

# 🔐 NATO Counter-Drone Operations in Belgium: Deep Technical Analysis for Cyber Security...

Author: Berend Watchus

Nov 12



In System Weakness by Berend Watchus

## The DJI Drone Ban: Comprehensive 2025 Update for Cybersecurity Professionals

Author: Berend Watchus includes free download in pdf

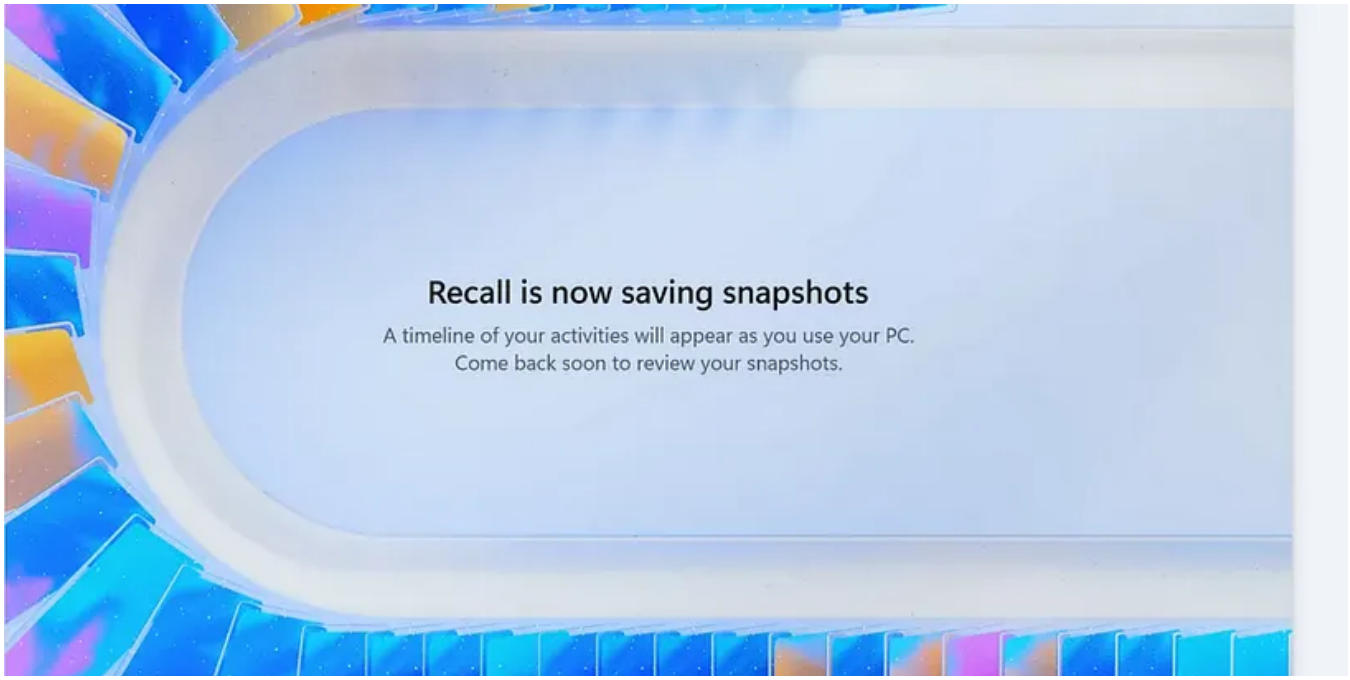
Nov 24

In System Weakness by Berend Watchus

## Whisper Leak: Your LLM Chats Are Not as Private as You Think (Even Encrypted!)

Author: Berend Watchus November 11, 2025 Blog article

Nov 7




In System Weakness by Berend Watchus

## Windows Recall: The VBS Enclave Paradox and the Evolution of Endpoint Espionage

Author: Berend Watchus November 11, 2025 Blog article

Nov 11 1

See all from Berend Watchus

## Recommended from Medium

Gabriel Odusanya | Cybersecurity

# Understanding the Modern Web Attack Surface (AppSec)

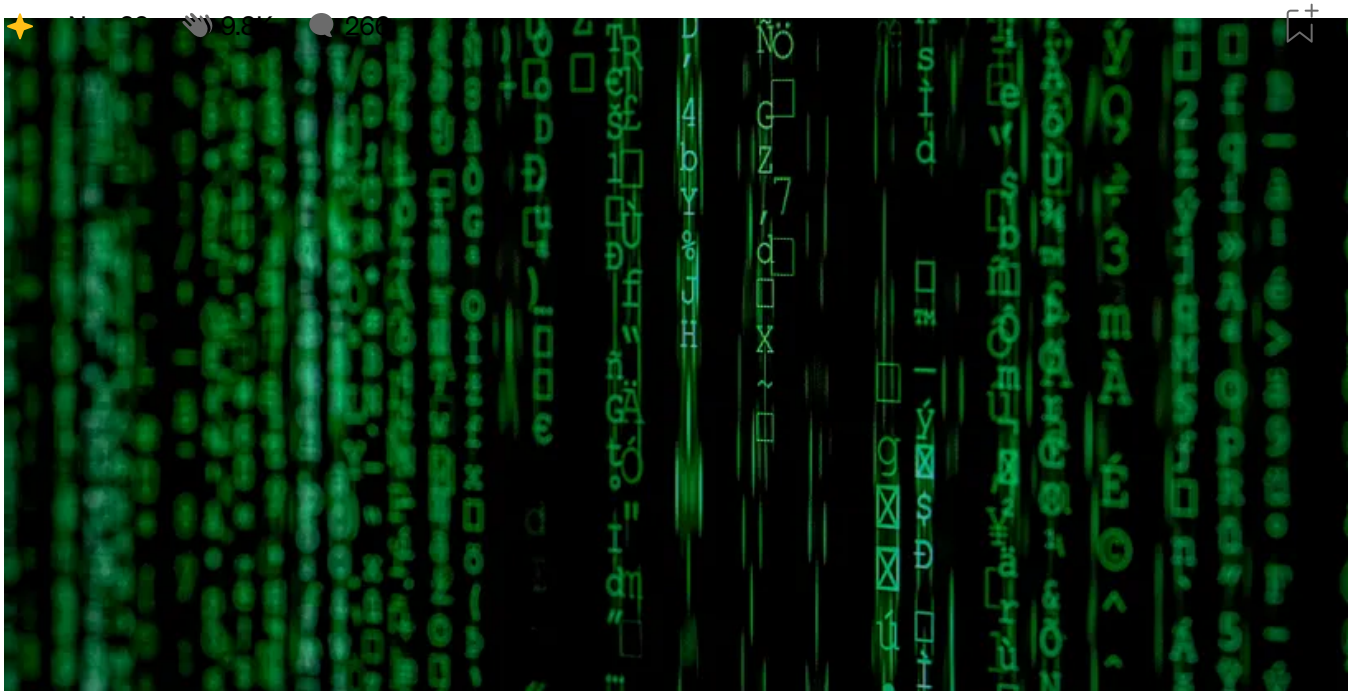
A practical breakdown based on hands-on reconnaissance and traffic analysis

5d ago

Will Lockett

## Peter Thiel Just Revealed How Utterly Screwed The Entire AI Industry Is

The AI bubble is bursting.

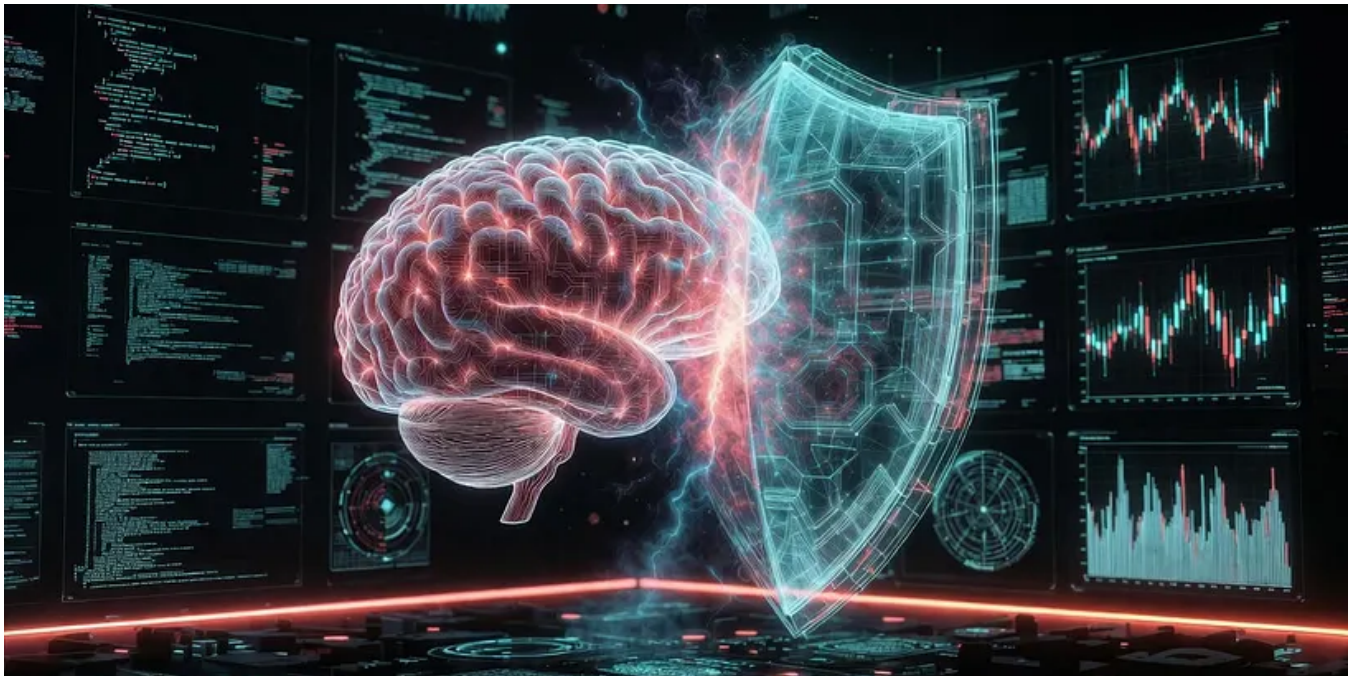


Very Lazy Tech

## Detecting Deepfake Phishing Calls: 10 Tools Every Cybersecurity Pro Should Master

What if you answered a call from the CEO, and it sounded exactly like them — voice, cadence, the whole deal — but it was a hacker on the...

5d ago 8 2

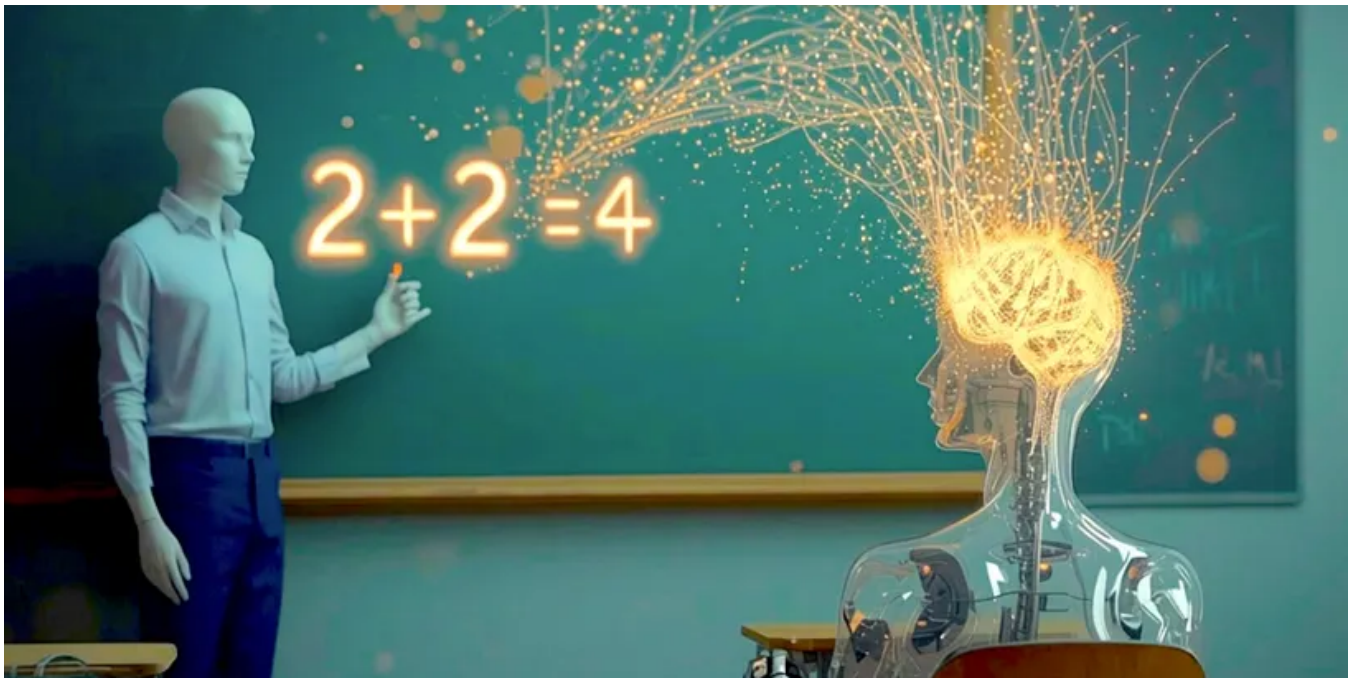


Maxwell Cross

# Weaponizing LLMs to Bypass Behavioral EDR

How Large Language Models Are Rewriting the Rules of Endpoint Detection Evasion

6d ago 11

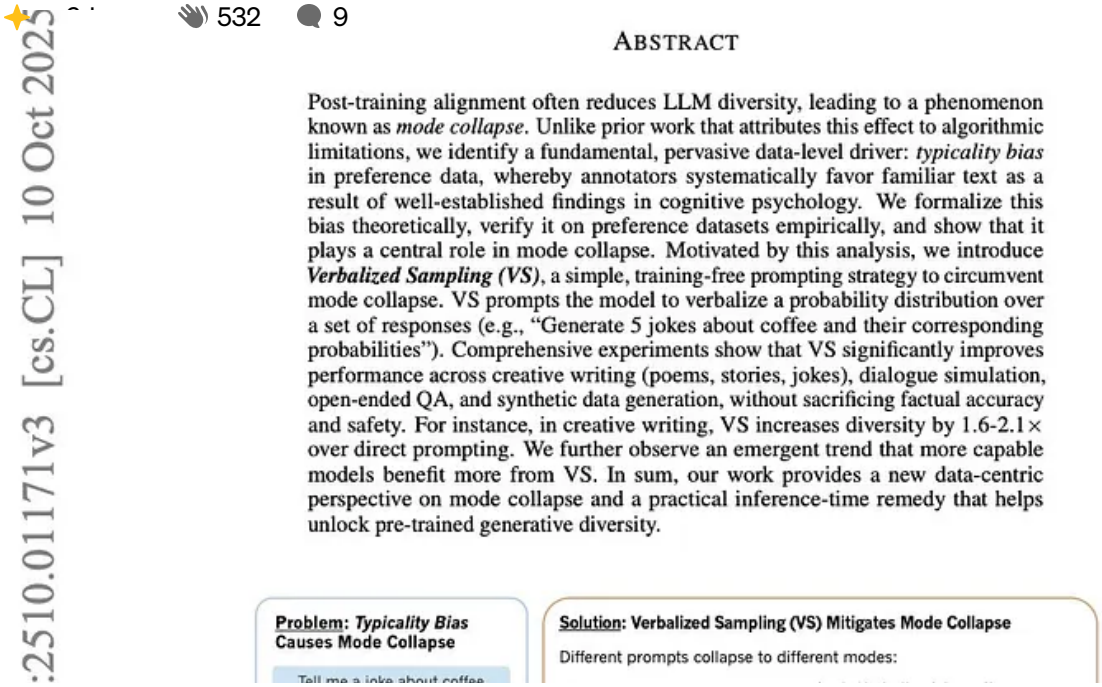


In AI Advances by Nikhil Anand

## I wasted years confused about how LLMs learn before learning this.

I wish someone had told me this when I started.

532 9



:2510.01171v3 [cs.CL] 10 Oct 2025

ABSTRACT

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental, pervasive data-level driver: *typicality bias* in preference data, whereby annotators systematically favor familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce ***Verbalized Sampling (VS)***, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., "Generate 5 jokes about coffee and their corresponding probabilities"). Comprehensive experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by 1.6-2.1× over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

**Problem: *Typicality Bias* Causes Mode Collapse**

**Solution: Verbalized Sampling (VS) Mitigates Mode Collapse**

Different prompts collapse to different modes:

In Generative AI by Adham Khaled

## Stanford Just Killed Prompt Engineering With 8 Words (And I Can't Believe It Worked)

ChatGPT keeps giving you the same boring response? This new technique unlocks 2× more creativity from ANY AI model — no training required...

Oct 20 19.3K 475

See more recommendations